INFO 6210
Data Management and Database
Final Project 1

I chose my domain as soccer. I scraped the dataset from Kaggle. https://www.kaggle.com/ankitasahni/exploring-soccer-analysis/data.

As it is known to us all, Soccer is a very popular sport in the world, great soccer clubs or players always have a huge quantity of fans, and a big event about soccer often aroused public attention and discussions on social media. So, it is a good domain to make analyses. In this project, I will model and gather data about my topic from Twitter.

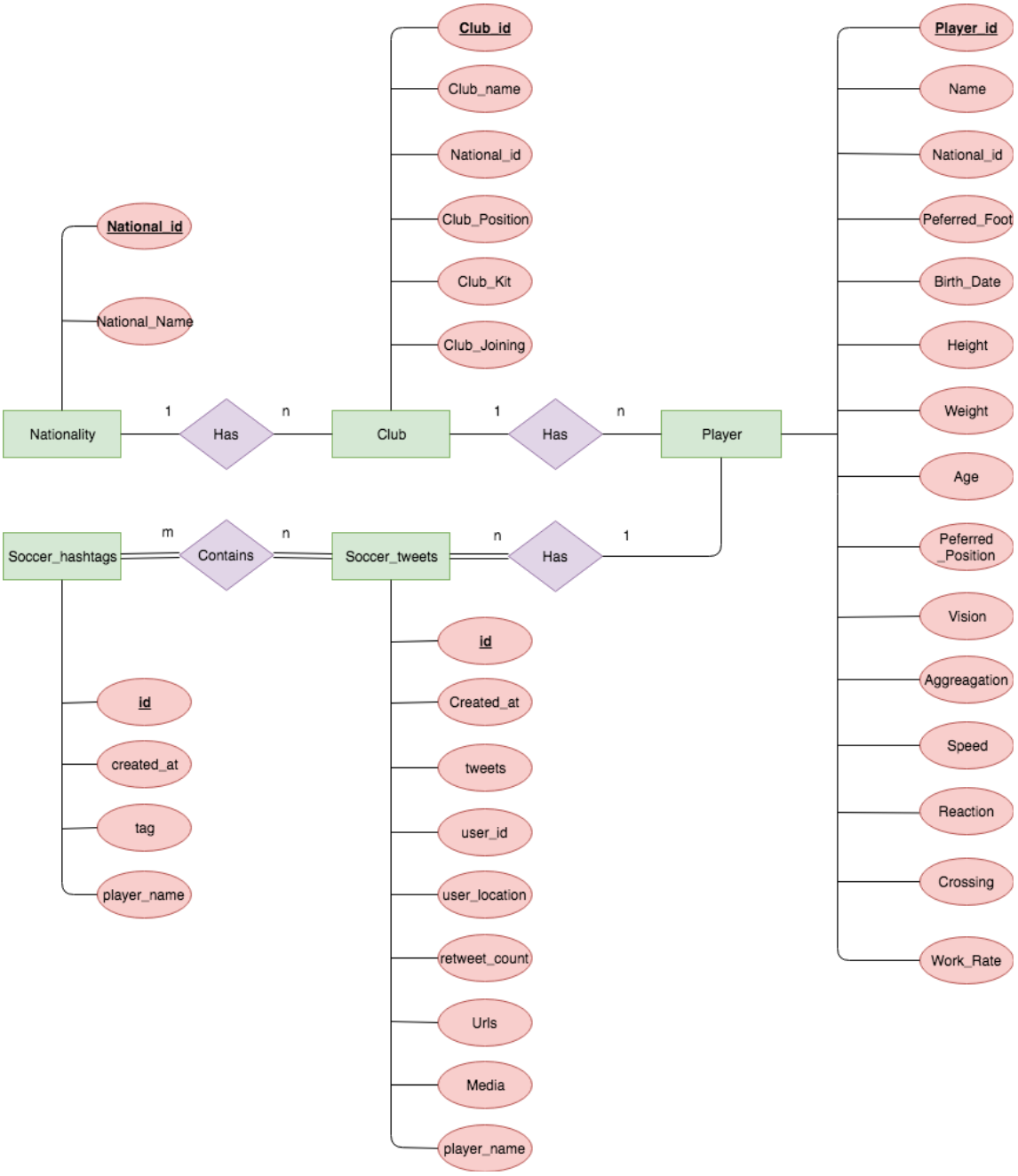
There are three parts in the project: soccer data, social media data and 3 tables including misspelled, synonyms and semantic.

Soccer data has 4 tables: soccer full data, nationality, club, player.

Social media data include 3 table: tweet full data, tweets and hashtags.

There are 5 entities in my project: Nationality, Club, Player, Soccer hashtags , Soccer tweets .

E-R Diagram:



Sample Data:

Player

index	Name	National_id	Height	Weight	Preffered_Foot	Birth_Date	Age	Preffered_Position	Work_Rate
0	Cristiano Ronaldo	a6	185 cm	80 kg	Right	02/05/1985	32	LW/ST	High / Low
1	Lionel Messi	a5	170 cm	72 kg	Left	06/24/1987	29	RW	Medium / Medium
2	Neymar	a3	174 cm	68 kg	Right	02/05/1992	25	LW	High / Medium
3	Luis Su rez	a11	182 cm	85 kg	Right	01/24/1987	30	ST	High / Medium
4	Manuel Neuer	a2	193 cm	92 kg	Right	03/27/1986	31	GK	Medium / Medium
5	De Gea	a1	193 cm	82 kg	Right	11/07/1990	26	GK	Medium / Medium
6	Robert Lewandowski	a12	185 cm	79 kg	Right	08/21/1988	28	ST	High / Medium
7	Gareth Bale	a24	183 cm	74 kg	Left	07/16/1989	27	RW	High / Medium
8	Zlatan Ibrahimovi	a21	195 cm	95 kg	Right	10/03/1981	35	ST	Medium / Low
9	Thibaut Courtois	a4	199 cm	91 kg	Left	05/11/1992	24	GK	Medium / Medium
10	Jrme Boateng	a2	192 cm	90 kg	Right	09/03/1988	28	CB	Medium / Medium

Nationality

index	National_id	Nationality_Name
0	a1	Spain
1	a2	Germany
2	a3	Brazil
3	a4	Belgium
4	a5	Argentina
5	a6	Portugal
6	a7	France
7	a8	Italy
8	a9	England
9	a10	Netherlands

Club

index	Club	National_id	Club_Position	Club_Kit	Club_Joining
0	Real Madrid	a6	LW	7	07/01/2009
1	FC Barcelona	a5	RW	10	07/01/2004
2	FC Barcelona	a3	LW	11	07/01/2013
3	FC Barcelona	a11	ST	9	07/11/2014
4	FC Bayern	a2	GK	1	07/01/2011
5	Manchester Utd	a1	GK	1	07/01/2011
6	FC Bayern	a12	ST	9	07/01/2014
7	Real Madrid	a24	RW	11	09/02/2013
8	Manchester Utd	a21	ST	9	07/01/2016
9	Chelsea	a4	GK	13	07/26/2011

Full data (the table that made up of “Player”, “Nationality” and “Club”)

index	"Name"	"National_id"	"Height"	"Weight"	"Preffered_Foot"	"Birth_Date"	"Age"	"Preffered_Position"	"Work_Rate"	"Aggression"	"Reacti
0	Cristiano Ronaldo	a6	185 cm	80 kg	Right	02/05/1985	32	LW/ST	High / Low	63	
1	Lionel Messi	a5	170 cm	72 kg	Left	06/24/1987	29	RW	Medium / Medium	48	
2	Neymar	a3	174 cm	68 kg	Right	02/05/1992	25	LW	High / Medium	56	
3	Luis Su rez	a11	182 cm	85 kg	Right	01/24/1987	30	ST	High / Medium	78	
4	Manuel Neuer	a2	193 cm	92 kg	Right	03/27/1986	31	GK	Medium / Medium	29	
5	De Gea	a1	193 cm	82 kg	Right	11/07/1990	26	GK	Medium / Medium	38	
6	Robert Lewandowski	a12	185 cm	79 kg	Right	08/21/1988	28	ST	High / Medium	80	
7	Gareth Bale	a24	183 cm	74 kg	Left	07/16/1989	27	RW	High / Medium	65	
8	Zlatan Ibrahimovi	a21	195 cm	95 kg	Right	10/03/1981	35	ST	Medium / Low	84	
9	Thibaut Courtois	a4	199 cm	91 kg	Left	05/11/1992	24	GK	Medium / Medium	23	

Data from social media sites :

Tweets

id	created_at	tweets	user_id	user_location	retweet_count
1	2018-04-17 17:50:07	RT @Sthabieeyr_77: Some people are talented #Lionel Mes…	834154403930238979	â- South Africa â- ,Ladysmithð	1
2	2018-04-17 03:42:41	RT @NoticiasVideos2: Atletico #Madrid #Vs #Barcelona #2…	936567941121888258	Miami, FL	2
4	2018-04-16 23:13:03	RT @NoticiasVideos2: Atletico #Madrid #Vs #Barcelona #2…	801715146590711808	*<null>*	2
5	2018-04-16 18:33:40	Atletico #Madrid #Vs #Barcelona #2017 1-2 | #Messi #Vs …	801715146590711808	*<null>*	2
6	2018-04-16 08:30:02	#Lionel Messi Vs #Cristiano Ronaldo https://t.co/xHwLiT…	984455546504990720	*<null>*	0
7	2018-04-16 05:06:14	Cristiano Ronaldo#CR7The Beast#LegendAnd The World…	862646148489773056	Islamabad, Pakistan	0
8	2018-04-15 16:10:56	#BabeKasihTau Membandingkan Reaksi Cristiano Ronaldo da…	2306171660	Indonesia	0
9	2018-04-15 14:41:13	#cristano ronaldo#Lionel messi	974246516050325504	snap:nicobj63	0
10	2018-04-13 20:56:58	RT @LibanoCule: #Lionel #Messi goal against Leganes , f…	970896576792195073	*<null>*	3

Hashtags

id	created_at	user_id	tag	player_name
1	2018-04-26 19:50:53	985538671288094720	Lionel	Lionel Messi
2	2018-04-26 19:41:10	2372357926	Messi	Lionel Messi
2	2018-04-26 19:41:10	2372357926	Lionel	Lionel Messi
3	2018-04-26 19:11:15	1101320424	Lionel	Lionel Messi
3	2018-04-26 19:11:15	1101320424	Messi	Lionel Messi
3	2018-04-26 19:11:15	1101320424	EuG	Lionel Messi
3	2018-04-26 19:11:15	1101320424	GOAT	Lionel Messi
3	2018-04-26 19:11:15	1101320424	FAZplus	Lionel Messi
4	2018-04-26 16:08:32	2306171660	BabeKasihTau	Lionel Messi
5	2018-04-26 15:52:07	717039627916484608	Lionel	Lionel Messi
5	2018-04-26 15:52:07	717039627916484608	Messi	Lionel Messi

Tweets full data (the table made up of “tweets” and “hashtags”)

id	created_at	tweets	user_id	user_location	tag
2642	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	stopmotion
2292	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	stopmotion
2292	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
1942	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
1592	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
192	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
542	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
1242	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
2642	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	neymar
892	2018-04-17 17:45:10	RT @NeymarJrSite: NÃ£o tire o olho do @neymarjr ...Que…	976187699857108994	Bloemfontein, South Africa	stopmotion

Code:

1. Loading soccer data

In [12]:
```
#connect to the database
import psycopg2
import pandas as pd
import sqlalchemy as sa
conn = psycopg2.connect(database='postgres',user='postgres', password='')
```

In [2]:
```
#use pandas to read csv table and load it to database
myfile1 = pd.read_csv('/Users/yizheliu/Desktop/DataBase/player.csv', encoding='latin1')
myfile2 = pd.read_csv('/Users/yizheliu/Desktop/DataBase/nationality.csv', encoding='latin1')
myfile3 = pd.read_csv('/Users/yizheliu/Desktop/DataBase/club.csv', encoding='latin1')
```

In [3]:
```
myfile1.duplicated().sum()
```

Out[3]:
```
0
```

In [4]:
```
myfile2.duplicated().sum()
```

Out[4]:
```
0
```

In [5]:
```
myfile3.duplicated().sum()
```

Out[5]:
```
0
```

In [14]:
```
con=sa.create_engine('postgresql://localhost/postgres')
```

In [7]:
```
#load table to database
myfile1.to_sql(name='soccer_player', if_exists='append', con=con)
myfile2.to_sql(name='soccer_nationality', if_exists='append', con=con)
myfile3.to_sql(name='soccer_club', if_exists='append', con=con)
```

In [8]:
```
myfile1.isnull().sum()
```

Out[8]:
```
index                0
Name                 0
National_id          0
Height               0
Weight               0
Preffered_Foot       0
Birth_Date           0
Age                  0
Preffered_Position    0
Work_Rate            0
Aggression           0
Reactions            0
Vision               0
Crossing             0
Speed                0
dtype: int64
```

In [9]:
```
myfile2.isnull().sum()
```

Out[9]:
```
National_id    0
Name           0
dtype: int64
```

In [10]:
```
myfile3.isnull().sum()
```

Out[10]:
```
index            0
Club             0
National_id      0
Club_Position    0
Club_Kit         0
Club_Joining     0
dtype: int64
```

2.Loading social media data

Scrape hashtags and tweets data from Twitter

In [1]:
```
import psycopg2
import pandas as pd
import tweepy

consumer_key = 'i2E80m2lW4azdS5FWylfckjK4'
consumer_secret = 'H9m8yJNaBloaMFISywaSmw7HkYM3hh3AfrOapLQHkWRWP5Ii9w'
access_token = '909086257308995584-FvP6rvrd1TSghqOFwruFKwY3rkuCc8Y'
access_secret = 'RUfqoSR7o40uJdpipfl3poV3eGs8uedZGxbBlB1WtChQx'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)
api = tweepy.API(auth, wait_on_rate_limit = True)
if (not api):
    print ("Problem connecting to API")
```

In [2]:
```
#Getting Tweets
import json

search_words = ['#Lionel Messi', '#Cristiano Ronaldo', '#Neymar', '#Wayne Rooney', '#Mario Balotelli', '#Andrea Pirlo',
search_name = ['Lionel Messi', 'Cristiano Ronaldo', 'Neymar', 'Wayne Rooney', 'Mario Balotelli', 'Andrea Pirlo', 'Garet
q = 'Soccer'
n = 105
from urllib.parse import unquote
search_results = api.search(q = q, count = n)
```

In [3]:
```
USER = "postgres"
PASSWORD = ""
DATABASE = "postgres"
```

In [5]:
```
import csv
tweets = []
n=1
for i in range (0,8):
        word=search_words[i]
        name=search_name[i]
        for tweet in api.search(q = word, count = 2000):
            for hashtag in tweet.entities.get('hashtags'):
                tweets.append([n,tweet.created_at, tweet.text, tweet.user.id,tweet.user.location, hashtag.get('text'),t
                print(tweet.created_at,tweet.text, tweet.user.id,tweet.user.location, hashtag, name)
            n=n+1

# csvfile = "tweets2.csv"
# with open(csvfile, "w") as output:
#     writer = csv.writer(output, lineterminator='\n')
#     for val in tweets:
#         writer.writerow([val])
```

```
2018-04-26 19:50:53 I'll love to have this nice tattoo #Lionel Messi https://t.co/bl4pJb3VCJ 987538671288094720 Niger
ia {'text': 'Lionel', 'indices': [35, 42]} Lionel Messi
2018-04-26 19:41:10 La rompió contra España estando 6-1. Y sin #Messi .. no busquen más un socio para #Lionel este vi
ene del mismo plan… https://t.co/jpazN40Xfi 2372357926 EL BOSQUE {'text': 'Messi', 'indices': [43, 49]} Lionel Messi
2018-04-26 19:41:10 La rompió contra España estando 6-1. Y sin #Messi .. no busquen más un socio para #Lionel este vi
ene del mismo plan… https://t.co/jpazN40Xfi 2372357926 EL BOSQUE {'text': 'Lionel', 'indices': [82, 89]} Lionel Messi
2018-04-26 19:11:15 #Lionel Andrés #Messi Cuccittini behält die Marke "Messi", erklärt das #EuG #GOAT 🐂⚽ https://
t.co/cQ9s86ZgDr #FAZplus via @faznet 1101320424 Köln I Frankfurt {'text': 'Lionel', 'indices': [0, 7]} Lionel Messi
2018-04-26 19:11:15 #Lionel Andrés #Messi Cuccittini behält die Marke "Messi", erklärt das #EuG #GOAT 🐂⚽ https://
t.co/cQ9s86ZgDr #FAZplus via @faznet 1101320424 Köln I Frankfurt {'text': 'Messi', 'indices': [15, 21]} Lionel Messi
2018-04-26 19:11:15 #Lionel Andrés #Messi Cuccittini behält die Marke "Messi", erklärt das #EuG #GOAT 🐂⚽ https://
t.co/cQ9s86ZgDr #FAZplus via @faznet 1101320424 Köln I Frankfurt {'text': 'EuG', 'indices': [71, 75]} Lionel Messi
2018-04-26 19:11:15 #Lionel Andrés #Messi Cuccittini behält die Marke "Messi", erklärt das #EuG #GOAT 🐂⚽ https://
t.co/cQ9s86ZgDr #FAZplus via @faznet 1101320424 Köln I Frankfurt {'text': 'GOAT', 'indices': [76, 81]} Lionel Messi
2018-04-26 19:11:15 #Lionel Andrés #Messi Cuccittini behält die Marke "Messi", erklärt das #EuG #GOAT 🐂⚽ https://
t.co/cQ9s86ZgDr #FAZplus via @faznet 1101320424 Köln I Frankfurt {'text': 'FAZplus', 'indices': [110, 118]} Lionel Me
ssi
2018-04-26 16:08:32 #BabeKasihTau Sepatu Emas Eropa 2017-2018: Mohamed Salah dan Lionel Messi Terdepan, Cristiano Ron
aldo Tak Masuk Hit… https://t.co/OstNiC3sYG 2306171660 Indonesia {'text': 'BabeKasihTau', 'indices': [0, 13]} Lionel
```

In [6]:
```
csvfile = "tweets.csv"
with open(csvfile, "w") as output:
    writer = csv.writer(output, lineterminator='\n')
    writer.writerows(tweets)
```

In [7]:
```
tweet = pd.read_csv('/Users/yizheliu/Desktop/DataBase/soccer_tweets.csv', encoding='latin1')
```

In [8]:
```
tweet.duplicated().sum()
```

Out[8]:
```
966
```

In [9]:
```
tweet2 = tweet.drop_duplicates()
```

In [10]:
```
tweet2.isnull().sum()
```

Out[10]:
```
id                  0
created_at          0
tweets              0
user_id             0
user_location      75
retweet_count       0
Urls                0
media             224
player_name         0
dtype: int64
```

In [15]:
```
tweet2.to_sql(name='soccer_tweets', if_exists='append', con=con)
```

In [16]:
```
hashtag = pd.read_csv('/Users/yizheliu/Desktop/DataBase/soccer_hashtags.csv', encoding='latin1')
```

In [17]:
```
hashtag.to_sql(name='soccer_hashtags', if_exists='append', con=con)
```

In [18]:
```
full_tweet = pd.read_csv('/Users/yizheliu/Desktop/DataBase/tweets.csv', encoding='latin1')
```

In [19]:
```
full_tweet.duplicated().sum()
```

Out[19]:
```
6
```

In [21]:
```
full_tweet.to_sql(name='full_tweet_data', if_exists='append', con=con)
```

3.Loading synonyms

In [93]:
```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/yizheliu/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
```

Out[93]:
```
True
```

In [94]:
```
from nltk.corpus import wordnet as wn
wn
```

Out[94]:
```
<WordNetCorpusReader in '/Users/yizheliu/nltk_data/corpora/wordnet'>
```

In [95]:
```
wn.synsets('car')
```

Out[95]:
```
[Synset('car.n.01'),
 Synset('car.n.02'),
 Synset('car.n.03'),
 Synset('car.n.04'),
 Synset('cable_car.n.01')]
```

In [96]:
```
hashtag1 = pd.read_csv('/Users/yizheliu/Desktop/DataBase/soccer_hashtags.csv', encoding='latin1',skipinitialspace=True)
hashtag1
```

...

In [98]:
```
x = hashtag1['tag']
```

In [99]:
```
x
```

...

In [101]:
```
#create a table called hashtag_synonyms to store hashtag synonmous data
%sql CREATE TABLE hashtag_synonyms(tag varchar, synonmous varchar)
```

```
Done.
```

Out[101]: []

In [102]:
```
#function, insert hashtag and corresponding synonyms to hashtag_synonyms table
def write_sql(tag, synonmous):
    db= psycopg2.connect(database="postgres", user="postgres", password="")
    cursor = db.cursor()
    insert_query = "INSERT INTO hashtag_synonyms(tag, synonmous) VALUES ( %s,%s)"
    cursor.execute(insert_query, (tag, synonmous))
    db.commit()
    cursor.close()
    db.close()
    return
```

In [103]:
```
#get synonyms of hashtags and write data to database
for hashtag in x:
    y = wn.synsets(hashtag)
    for syn in y:
        for l in syn.lemmas():
            write_sql(hashtag, l.name())
            print(hashtag, y)
#     for t in y:
#         c = wn.synsets(t)
```

...

In [104]:
```
#select distinct data from hashtags_synonyms, insert these data to another table called hashtag_synonyms, this step
# to delete repeat rows
%sql select distinct * into hashtags_synonyms from hashtag_synonyms
```

```
791 rows affected.
```

Out[104]: []

In [105]:
```
#drop original table
%sql DROP TABLE hashtag_synonyms
```

...

In [106]:
```
#show data from new table without repeat information
%sql SELECT * FROM hashtags_synonyms
```

...

4.Create misspelled words table

In [107]:
```
#create a table to store misspelled words
%sql create table hashtag_misspelled(tag varchar, chech varchar, missepelled varchar)
```

```
Done.
```

Out[107]: []

In [109]:
```
#import enchant library to check if this word is misspelled and get the suggest correct word
import enchant
d = enchant.Dict("en_US")
```

In [110]:
```
d.check('hello')
d.suggest('hello')[0]
```

Out[110]: 'hello'

In [111]:
```
#function, insert misspelled words data to database
def write_sql1(tag,chech,missepelled):
    db= psycopg2.connect(database="postgres", user="postgres", password="")
    cursor = db.cursor()
    insert_query = "INSERT INTO hashtag_misspelled(tag,chech,missepelled) VALUES (%s,%s,%s)"
    cursor.execute(insert_query, (tag,chech,missepelled))
    db.commit()
    cursor.close()
    db.close()
    return
```

In [112]:
```
#get misspelled data and insert it to database. including hashtags, if correct, suggest words
for i in range(0,641):
    y = d.check(x[i])
    z = d.suggest(x[i])
    for t in z:
        write_sql1(x[i], y, t)
        print(x[i], y, t)
```

```
Lionel True Linoel
Lionel True Leonel
Lionel True Lion el
Lionel True Lionel
Messi False Nessi
Messi False Tessi
Messi False Mess
```

In [113]:
```
#delete repeat rows, select distinct data to new table, drop old one
%sql select distinct * into misspelled_hashtags from hashtag_misspelled
```

```
946 rows affected.
```

Out[113]: []

In [114]:
```
%sql DROP TABLE hashtag_misspelled
```

```
Done.
```

Out[114]: []

In [115]:
```
#show misspelled data from new table
%sql SELECT * FROM misspelled_hashtags
```

...

5. create Semantic Information table

In [116]:
```
##create a table to store hashtag_semantic_category data, including 2 fields, hashtag and category
%sql create table hashtag_semantic_category(tag varchar, category varchar)
```

```
Done.
```

Out[116]: []

In [117]:
```
#function, insert getting data to database
def write_sql2(tag,category):
    db= psycopg2.connect(database="postgres", user="postgres", password="")
    cursor = db.cursor()
    insert_query = "INSERT INTO hashtag_semantic_category(tag,category) VALUES (%s,%s)"
    cursor.execute(insert_query, (tag,category))
    db.commit()
    cursor.close()
    db.close()
    return
```

In [118]:
```
#loop, get hashtag category data by using nltk library and load it to table hashtag_semantic_category
for hashtag in x:
    y = wn.synsets(hashtag)
    for syn in y:
        for l in syn.topic_domains():
            for n in l.lemmas():
                write_sql2(hashtag, n.name())
                print(hashtag, n.name())
```

...

In [119]:
```
#delete repeat rows, select distinct data to new table, drop old one
%sql select distinct * into hashtag_semantic_categories from hashtag_semantic_category
```

```
78 rows affected.
```

Out[119]: []

In [120]:
```
%sql DROP TABLE hashtag_semantic_category
```

```
Done.
```

Out[120]: []

In [121]:
```
#show hashtag category data from new table
%sql SELECT * FROM hashtag_semantic_categories
```

...

Questions

In [23]:
```
%load_ext sql
```

In [24]:
```
%sql postgresql://postgres@127.0.0.1/postgres
```

Out[24]: 'Connected: postgres@postgres'

1.What are people saying about me (somebody)?

We can find all the tags related to Lionel Messi(somebody) in the soccer_hashtags table. Also, we can get both the tweets and hashtags after making joins between soccer_hashtags and soccer_tweets.

In [25]:
```
%sql SELECT tag FROM soccer_hashtags WHERE player_name LIKE 'Lionel Messi'
```

```
259 rows affected.
```

Out[25]:

tag
Lionel
Messi
Lionel
Lionel
Messi
EuG
GOAT
FAZplus
BabeKasihTau
Lionel
Messi

In [126]:
```
%sql SELECT soccer_hashtags.tag, soccer_tweets.tweets FROM soccer_hashtags JOIN soccer_tweets on soccer_tweets.id = soc
WHERE soccer_tweets.player_name LIKE '%Cristiano Ronaldo%'
```

```
374 rows affected.
```

Out[126]:

tag	tweets
realmadrid	RT @defcentral: El plan de Cristiano Ronaldo para llegar a tope al choque de vuelta ante el Bayern: https://t.co/qTjk63NrvT #realmadrid #Crâ¦
Neyma	RT @Ntswaki70509969: Who is the best player?? Rt for #Neyma OR like for #Cristiano Ronaldo https://t.co/bg6455Bhhv
Cristiano	RT @Ntswaki70509969: Who is the best player?? Rt for #Neyma OR like for #Cristiano Ronaldo https://t.co/bg6455Bhhv
CristianoRonaldo	RT @FootTheBall: Â©ï¸Â®ï¸7ï¸â£ fans: He's the BEST NOT Â©ï¸Â®ï¸7ï¸â£ fans: ðð #CristianoRonaldo #Cristiano #CR7 #Ronaldo #RealMadrid #BayernReal #Fooâ¦

2.How viral are my posts?

I choose to use the retweet count to show one's posts impact.

In [29]:
```
%sql select user_id, retweet_count from soccer_tweets order by retweet_count desc
```

704594716474519552	2
103355620	2
2889076047	2
704594716474519552	2
3242927754	2
3242927754	2
589238345	2
704594716474519552	2
4733730982	2
51847064	2
988000149858406402	2
2280450452	2
2399322038	2

3.How much influence to my posts have?

We can know the influence by ordering the retweet count and favorite count, however, twitter official website doesn't give the access to get favorite count. So, after finding the user who has the most retweets, we get his/her statuses_count, friends_count and followers_count.

In [127]:
```
%sql select user_id, retweet_count from soccer_tweets order by retweet_count desc
```

```
264 rows affected.
```

Out[127]:

user_id	retweet_count
868210071788040193	40
982266942441467904	40
781931679816572929	40
761167991480410112	40
368588734	40
846056660086984704	30
269508306	30
2678504636	30
2257539278	30
212470952	30
66666126	30

In [178]:
```
u = api.get_user(868210071788040193)
print(u.screen_name)
```

```
Hassan_dzzz
```

In [179]:
```
print(str(u.statuses_count), str(u.friends_count), str(u.followers_count))
```

```
156 357 11
```

4.What posts are like mine?

For this question, we can get tweets which contain the same tag. Here are two example with tag "Messi" and tag "Cristiano".

In [134]:
```
%sql SELECT soccer_hashtags.tag, soccer_tweets.tweets \
FROM soccer_hashtags JOIN soccer_tweets on soccer_tweets.id = soccer_hashtags.id WHERE soccer_hashtags.tag = 'Messi'
```

```
29 rows affected.
```

Out[134]:

tag	tweets
Messi	La rompiÃ³ contra EspaÃ±a estando 6-1. Y sin #Messi .. no busquen mÃ¡s un socio para #Lionel este viene del mismo planâ¦ https://t.co/jpazN40Xfi
Messi	#Lionel AndrÃ©s #Messi Cuccittini behÃ¤lt die Marke "Messi", erklÃ¤rt das #EuG #GOAT ðâ½ï¸ https://t.co/cQ9s86ZgDr #FAZplus via @faznet
Messi	Lionel Messi vince anche al Tribunale Ue: puÃ² registrare il suo marchio - #Lionel #Messi #vince #anche https://t.co/24VoYgIoxt
Messi	Lionel Messi vince anche al Tribunale Ue: puÃ² registrare il suo marchio - #Lionel #Messi #vince #anche https://t.co/qPbAzaGrG3
Messi	Lionel Messi puÃ² registrare marchio per abbigliamento - #Lionel #Messi #registrare #marchio https://t.co/iCH9bnyfKX
Messi	Messi scores in EU court battle to trademark name #LionelMessi #Messi #Lionel #EU #Trademark #MESSI #Massiâ¦ https://t.co/H9zq3dRvPk
Messi	#Foto #polÃªmica #sugere que #Lionel #Messi tem #seis #dedos no pÃ© #direito; #confira https://t.co/0Ow0MBlu3x
Messi	Gol #Messi final del Mundialito de Clubs #Fcbarcelona 2-1 #Estudiantes online https://t.co/RQjVjrykLS #BARCAâ¦ https://t.co/ENX2cMmlAn
Messi	RT @NoticiasVideos2: #Goles #De #Messi visto #Desde #La tribuna online https://t.co/UrLQTbdcvx #ESTADIO #FcBarcelona #Fcbarcelona #Grada #Hâ¦

In [135]:
```
%sql SELECT soccer_hashtags.tag, soccer_tweets.tweets \
FROM soccer_hashtags JOIN soccer_tweets on soccer_tweets.id = soccer_hashtags.id WHERE soccer_hashtags.tag = 'Cristiano
```

```
29 rows affected.
```

Out[135]:

tag	tweets
Cristiano	Tampil Hebat, Rating Salah di FIFA 18 Melonjak Drastis! ---------- #Cristiano Ronaldo #FIFA 18 #Lionel Messiâ¦ https://t.co/2wrEdC2I3Q
Cristiano	RT @Ntswaki70509969: Who is the best player?? Rt for #Neyma OR like for #Cristiano Ronaldo https://t.co/bg6455Bhhv
Cristiano	RT @FootTheBall: Â©ï¸Â®ï¸7ï¸â£ fans: He's the BEST NOT Â©ï¸Â®ï¸7ï¸â£ fans: ðð #CristianoRonaldo #Cristiano #CR7 #Ronaldo #RealMadrid #BayernReal #Fooâ¦
Cristiano	RT @FootTheBall: ð¨ is coming.....next season #ChampionsLeague #UCL #Cristiano #Ronaldo #Casillas #FootTheBall https://t.co/vD4iOQf7aO
	RT @FootTheBall: Lucky Bashford! Â

5.What users post like me?

We can find users who are smilar to each other by their shared tags. The following are two examples of the tag "Messi" and the tag "Cristiano".

In [132]:
```
%sql SELECT soccer_tweets.user_id, soccer_hashtags.tag \
FROM soccer_hashtags JOIN soccer_tweets on soccer_tweets.id = soccer_hashtags.id WHERE soccer_hashtags.tag = 'Messi'
```

```
29 rows affected.
```

Out[132]:

user_id	tag
2372357926	Messi
1101320424	Messi
717039627916484608	Messi
231726084	Messi
717039627916484608	Messi
290417646	Messi
207333431	Messi
801715146590711808	Messi
920239369549746177	Messi
801715146590711808	Messi

In [133]:
```
%sql SELECT soccer_tweets.user_id, soccer_hashtags.tag \
FROM soccer_hashtags JOIN soccer_tweets on soccer_tweets.id = soccer_hashtags.id WHERE soccer_hashtags.tag = 'Cristiano'
```

```
29 rows affected.
```

Out[133]:

user_id	tag
847142958	Cristiano
989569276112769024	Cristiano
3242927754	Cristiano
3242927754	Cristiano

6.Who should I be following?

We always follow people who are similar to us and also have a high quality. So, I get his description to see his interest. Besides, I would like to get the user's followers, average tweets per day and other data showing his/her quality. Morever,I can download all his/her tweets, hashtags and mentions to see his posts'quality and compare his/her interests with me.

In [136]:
```
user = api.get_user('WeAreMessi')
print(user.name, user.screen_name)
```

```
Leo Messi 🔟 WeAreMessi
```

In [137]:
```
print(str(user.statuses_count), str(user.friends_count), str(user.followers_count))
```

```
29315 476 107776
```

In [138]:
```
print(user.description)
```

```
#1 Fan Club of Lionel Messi on Twitter! ⚽ || Instagram: We.Are.Messi 📷 || Facebook & Snapchat: WeAreMessi 👻
```

In [139]:
```
from datetime import datetime, date, time, timedelta
tweets = user.statuses_count
account_created_date = user.created_at
delta = datetime.utcnow() - account_created_date
account_age_days = delta.days
print("Account age (in days): " + str(account_age_days))
if account_age_days > 0:
    print("Average tweets per day: " + "%.2f"%(float(tweets)/float(account_age_days)))
```

```
Account age (in days): 1295
Average tweets per day: 22.64
```

In [140]:
```
from tweepy import Cursor
hashtags = []
mentions = []
tweet_count2 = 0
end_date = datetime.utcnow() - timedelta(days=30)
for status in Cursor(api.user_timeline, id='WeAreMessi').items():
    tweet_count2 += 1
    if hasattr(status, "entities"):
        entities = status.entities
        if "hashtags" in entities:
            for ent in entities["hashtags"]:
                if ent is not None:
                    if "text" in ent:
                        hashtag = ent["text"]
                if hashtag is not None:
                    hashtags.append(hashtag)
        if "user_mentions" in entities:
            for ent in entities["user_mentions"]:
                if ent is not None:
                    if "screen_name" in ent:
                        name = ent["screen_name"]
                if name is not None:
                    mentions.append(name)
        if status.created_at < end_date:
            break
```

In [141]:
```
hashtags
```

...

In [142]:
```
mentions
```

In [143]:
```
import csv
res1=hashtags
csvfile = "Q6_hashtags.csv"
with open(csvfile, "w") as output:
    writer = csv.writer(output, lineterminator='\n')
    for val in res1:
        writer.writerow([val])
```

In [144]:
```
res2=mentions
csvfile = "Q6_mentions.csv"
with open(csvfile, "w") as output:
    writer = csv.writer(output, lineterminator='\n')
    for val in res2:
        writer.writerow([val])
```

In [160]:
```
newfile1 = pd.read_csv('/Users/yizheliu/Desktop/DataBase/Q6_hashtags.csv', encoding='latin1')
newfile2 = pd.read_csv('/Users/yizheliu/Desktop/DataBase/Q6_mentions.csv', encoding='latin1')
```

In [161]:
```
newfile1.to_sql(name='Q6_hashtags', if_exists='append', con=con)
newfile2.to_sql(name='Q6_mentions', if_exists='append', con=con)
```

In [153]:
```
df_r = pd.read_sql('''select hashtags, count(hashtags) as sum from "Q6_hashtags" group by hashtags order by sum desc'''
                   con = con)
```

In [170]:
```
df_r.head(10)
```

Out[170]:

	hashtags	sum
0	WeAreMessi	136
1	Messi	120
2	fcblive	26
3	CopaBarÃ§a	22
4	BarÃ§aRoma	16

In [167]:
```
df_s = pd.read_sql('''select mentions, count(mentions) as sum from "Q6_mentions" group by mentions order by sum desc'''
                   con = con)
```

In [169]:
```
df_s.head(10)
```

Out[169]:

	mentions	sum
0	WeAreMessi	275
1	SevillaFC	9
2	OfficialASRoma	8
3	FCBarcelona	7
4	brfootball	6
5	Marwan_elzahar	6
6	sandraD10S	5
7	Suyeb99	5
8	valenciacf	4
9	Lionelmypurpose	4

7.What topics are trending in my domain?

We can order the amount of appearances of all of the tags over a specified time. Morever, we can compare the change of hashtags between two different time to predict future trending.

In [47]:
```
%sql select created_at, player_name, retweet_count from soccer_tweets order by created_at desc
```

```
264 rows affected.
```

Out[47]:

created_at	player_name	retweet_count
2018-04-26 21:12:09	Neymar	19
2018-04-26 21:02:47	Cristiano Ronaldo	1
2018-04-26 21:01:59	Neymar	1
2018-04-26 21:00:50	Neymar	1
2018-04-26 20:47:14	Neymar	2
2018-04-26 20:42:12	Neymar	40
2018-04-26 20:42:12	Neymar	40
2018-04-26 20:27:46	Neymar	1
2018-04-26 20:23:25	Neymar	2
2018-04-26 20:22:26	Neymar	2
2018-04-26 20:19:51	Neymar	40

In [48]:
```
%sql SELECT tag ,count(*) AS cout FROM soccer_hashtags WHERE created_at BETWEEN '2018-04-24 00:00:00' \
and  '2018-04-25 23:59:59' GROUP BY tag ORDER BY COUT DESC
```

```
150 rows affected.
```

Out[48]:

tag	cout
BayernMunich	40
BayernRealMadrid	38
UCL	37
Marcelo	27
laliga	16
realmadrid	16
harryfan	15
spain	15
welsh	15
Wales	15
garethbale11	15

In [49]:
```
%sql SELECT tag ,count(*) AS cout FROM soccer_hashtags WHERE created_at BETWEEN '2018-04-25 00:00:00' \
and  '2018-04-26 23:59:59' GROUP BY tag ORDER BY COUT DESC
```

```
269 rows affected.
```

Out[49]:

tag	cout
Neymar	60
UCL	43
BayernMunich	40
BayernRealMadrid	38
Cristiano	29
Marcelo	27

8.What keywords/ hashtags should I add to my post?

By counting the sum of retweet count of the same tag and ordering it, we can know that the post contains which tags is popular in a specific domain. These tags are what we would like to add to our posts.

In [62]:
```
%sql SELECT soccer_hashtags.tag, sum(soccer_tweets.retweet_count) as s \
FROM soccer_tweets JOIN soccer_hashtags ON soccer_hashtags.id = soccer_tweets.id \
WHERE soccer_tweets.player_name = 'Cristiano Ronaldo' \
GROUP by soccer_hashtags.tag \
ORDER BY s DESC;
```

```
88 rows affected.
```

Out[62]:

tag	s
UCL	897
BayernMunich	884
BayernRealMadrid	882
Marcelo	810
Asensio	72
Cristiano	43
Ronaldo	28
ChampionsLeague	21
Bycyclekick	12
CristianoJr	12
RealMadrid	10

9.Should I follow somebody back?

The main factor to determine if we should follow somebody back is his/her quality. So, I will get the user's statuses count, followers count, average tweets per day and so on. Also, by downloading the user's tweets, hashtags and mentions,I can know if the user's interests are like mine.

In [66]:
```
user = api.get_user('BarcaWorldwide')
print(user.name, user.screen_name)
```

```
Barcelona Worldwide BarcaWorldwide
```

In [67]:
```
print(str(user.statuses_count), str(user.friends_count), str(user.followers_count))
```

```
49994 40 53736
```

In [68]:
```
print(user.description)
```

```
An account dedicated to the best club in the world, FC Barcelona and to its fans around the globe🌍
```

In [69]:
```
from datetime import datetime, date, time, timedelta
tweets = user.statuses_count
account_created_date = user.created_at
delta = datetime.utcnow() - account_created_date
account_age_days = delta.days
print("Account age (in days): " + str(account_age_days))
if account_age_days > 0:
    print("Average tweets per day: " + "%.2f"%(float(tweets)/float(account_age_days)))
```

```
Account age (in days): 1246
Average tweets per day: 40.12
```

In [70]:
```
from tweepy import Cursor
hashtags = []
mentions = []
tweet_count2 = 0
end_date = datetime.utcnow() - timedelta(days=30)
for status in Cursor(api.user_timeline, id='BarcaWorldwide').items():
    tweet_count2 += 1
    if hasattr(status, "entities"):
        entities = status.entities
        if "hashtags" in entities:
            for ent in entities["hashtags"]:
                if ent is not None:
                    if "text" in ent:
                        hashtag = ent["text"]
                if hashtag is not None:
                    hashtags.append(hashtag)
        if "user_mentions" in entities:
            for ent in entities["user_mentions"]:
                if ent is not None:
                    if "screen_name" in ent:
                        name = ent["screen_name"]
                if name is not None:
                    mentions.append(name)
        if status.created_at < end_date:
            break
```

In [71]:
```
hashtags
```

Out[71]:
```
['ForçaBarça',
 'GraciasIniesta',
 'LEGEND',
 'SeQueda',
 'Messiesque',
 'RoadToKyiv',
 'GodOfFootball',
 'GOAT',
 'ForçaBarça']
```

10.What is the best time to post?

To know the best time to post, we can get the time of post which has most retweets for the relevant tags in the post.

In [174]:
```
%sql SELECT created_at,tag,tweets, retweet_count from full_tweet_data where player_name = 'Neymar'\
order by retweet_count desc
```

```
346 rows affected.
```

Out[174]:

created_at	tag	tweets	retweet_count
2018-04-26 20:02:58	Neymar	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40
2018-04-26 20:02:56	Locelso	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40
2018-04-26 20:02:56	draxler	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40
2018-04-26 20:02:56	PSGASM	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40
2018-04-26 20:02:56	dimaria	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40
2018-04-26 20:10:51	psg	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40
2018-04-26 20:10:51	PSGASM	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	40

In [177]:
```
%sql SELECT created_at,tag,tweets, retweet_count from full_tweet_data where player_name = 'Cristiano Ronaldo'\
order by retweet_count desc
```

```
374 rows affected.
```

Out[177]:

created_at	tag	tweets	retweet_count
2018-04-25 19:46:24	UCL	RT @FootballAdictt: Marcelo Goal HD - Bayern Munich1-1Real Madrid 25.04.2018 #Marcelo #BayernRealMadrid #BayernMunich #UCL #ChampionsLeaguâ¦	30
		RT @FootballAdictt: Marcelo Goal HD - Bayern Munich1-1Real Madrid 25.04.2018	

11.Should I add and picture or url to my post?

We can compare the retweets of the posts with url or picture to the retweets of the posts without url and picures.

In [91]:
```
%sql SELECT soccer_tweets.* from soccer_tweets where "media" IS \
                        NULL AND "Urls" IS NULL order by retweet_count DESC
```

```
81 rows affected.
```

Out[91]:

index	id	created_at	tweets	user_id	user_location	retweet_count	Urls	media	player_name
672	170	2018-04-26 20:10:51	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	868210071788040193	France	40	None	None	Neymar
689	173	2018-04-26 20:02:56	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	982266942441467904	None	40	None	None	Neymar
681	172	2018-04-26 20:02:58	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	368588734	Instagram - Snap: jdsjulian	40	None	None	Neymar
648	165	2018-04-26 20:42:12	RT @Aanaiis41: CHAMPIONS ð PARIS SG @PSG_inside @Ligue1Conforama #psg #PSGASM #dimaria #cavani #Locelso #draxler #mbappe #Neymar #championâ¦	761167991480410112	Brisbane, Queensland	40	None	None	Neymar

In [92]:
```
%sql SELECT soccer_tweets.* from soccer_tweets where "media" IS \
                        NOT NULL OR "Urls" IS NOT NULL order by retweet_count DESC
```

```
183 rows affected.
```

Out[92]:

index	id	created_at	tweets	user_id	user_location	retweet_count	
620	158	2018-04-25 19:35:02	Marcelo Goal HD - Bayern Munich1-1Real Madrid 25.04.2018 #Marcelo #BayernRealMadrid #BayernMunich #UCLâ https://t.co/FNTK7wmi4o	927207083539251200	Ahmadabad City, India	30	[{'url': 'https://t.co/FNTK7wmi4o', 'expa 'https://twitter.com/i/web/status/98922629523 'display_url': 'twitter.com/i/web/status/9â', 'indices': [

	620	158	2018-04-25 19:35:02	#Marcelo #BayernRealMadrid #BayernMunich #UCLâ¦ https://t.co/FNTK7wmi4o	927207083539251200	Ahmadabad City, India	30	[{ url': 'https://t.co/FNTK7wmi4o', 'expa 'https://twitter.com/i/web/status/98922629523 'display_url': 'twitter.com/i/web/status/9â¦', 'indices': [
	927	240	2018-04-26 13:49:53	RT @CanalSupporters: Nickelodeon lance un dessin animÃ© inspirÃ© par Neymar https://t.co/hu0WG6f4YH #PSG #teamPSG #Nickelodeon #Neymar https:â¦	2597580620	None	12	[{'url': 'https://t.co/hu0WG6f4YH', 'expanded_url': 'ht supporters.com/2018/04/nickelodeon-lance-un-dess inspire-par-neymar/', 'display_u supporters.com/2018/04/nickelâ¦', 'indices'
	968	248	2018-04-26 13:17:36	RT @CanalSupporters: Nickelodeon lance un dessin animÃ© inspirÃ© par Neymar https://t.co/hu0WG6f4YH #PSG #teamPSG #Nickelodeon #Neymar https:â¦	1544229428	Paris	12	[{'url': 'https://t.co/hu0WG6f4YH', 'expanded_url': 'ht supporters.com/2018/04/nickelodeon-lance-un-dess inspire-par-neymar/', 'display_u supporters.com/2018/04/nickelâ¦', 'indices'
				RT @CanalSupporters: Nickelodeon				

12.What's my reach?

> We can count a user's retweet_count, favriate_count and follows_count.

In [180]:
```
u = api.get_user('imessi')
print(u.id, str(u.followers_count))
```

```
50600749 295378
```

In [181]:
```
tweet = api.user_timeline(id = 50600749, count = 1)[0]
```

In [182]:
```
print(tweet.text, tweet.retweet_count, tweet.favorite_count)
```

```
Early christmas gift for you all:
Messi vs RealMadrid
https://t.co/rzhThEaV7Q 32 410
```